Ninth Conference of the European Chapter of the Association for Computational Linguistics

8–12 June 1999
University of Bergen
Bergen, Norway

Published by the Association for Computational Linguistics

The conference was sponsored by

LINGSOFT
University of Bergen (Humanities Faculty)
Bergen University Fund
Norwegian Ministry of Education, Research and Church affairs

Nordisk Språkteknologi AS (Nordic Language Technology)
Ministry of Culture

Order copies of this and other ACL publications from:

Morgan Kaufmann Publishers
340 Pine Street, 6th floor
San Francisco, CA 94104, USA
+1 800 745 7323

Preface

This volume contains the papers prepared for the Ninth Conference of the European Chapter of the Association for Computational Linguistics, held 8–12 June in Bergen, Norway.

The conference programme features invited talks, tutorials, submitted papers in both main and student sessions and a poster and demo session.

The main session Programme Committee received 99 abstracts from 20 countries around the world. Every paper was reviewed by at least three reviewers via a web-based interface which preserved anonymity. The fairness of the process is we think demonstrated by the broad spread of institutions and countries represented in the papers which were accepted and are printed here: 29 papers from 12 countries, with no country having more than 6 papers. The topics of the accepted papers cover a wide range of topics, and taken together we think they present an exciting and up-to-date sample of the best work in our field at the present time.

Our Area Chairs worked extremely hard in recruiting the Programme Committee, allocating papers to reviewers and encouraging the completion of reviews on time, and then joined the Programme Chairs for an intensive all-day meeting in Edinburgh where the final decisions were made. We are extremely grateful to both the Area Chairs and the Programme Committee for their hard work.

We used a new web-based approach to managing the paper submission and reviewing process this year, with authors registering an intention to submit electronically, and web-based reviewing processes, while at the same time retaining paper submission of the manuscripts themselves. We've had a lot of positive feedback about the reviewing system from the Programme Committee. We hope the system worked for authors as well.

Pulling together a meeting on this scale is a major effort, and thanks are due to many people for making it happen: Bruce Croft and Wolfgang Wahlster for the invited talks; Walter Daelemans for organising the tutorials and Robert Dale, Ronen Feldman, Adam Kilgarriff, Adwait Ratnaparkhi, Ehud Reiter and Michael Rundell for preparing and delivering them; the EACL President (Donia Scott) and past President (John Nerbonne) for their helpful suggestions; Gertjan van Noord for organising the post-conference workshops and Jonas Kuhn and Avro Voutilainen for organising the student sessions.

We are also very grateful to the Local Organisation Committee Chair, Koenraad de Smedt, for his patient cooperation with us over many issues, large and small, as well as his stirling efforts in making sure we would actually *have* a conference at which these papers could be presented.

We also are indebted to Robert Inder made his web-based registration and reviewing system available to us, and Richard Tobin helped us adapt it to our needs.

Local help in organising the Programme Committee meeting; with incoming and outgoing mail and with producing these Proceedings was provided by Janet Forbes and Margaret McMillan, to whom many thanks.

Henry S. Thompson, Alex Lascarides
Programme Committee Chairs

Programme Committee (Main Session)

Programme Chair:
Henry S. Thompson (University of Edinburgh)

Programme Co-Chair:
Alex Lascarides (University of Edinburgh)

Area Chairs:

John Carroll (University of Sussex)
Jan Hajic (Charles University)
Paul Taylor (University of Edinburgh)
Rob Gaizauskas (Sheffield University)
Mats Rooth (University of Stuttgart)
Lyn Walker (AT&T Bell Labs)

Programme Committee Members:

Ann Abeille (University of Paris 7)
Elisabeth Andre (DFKI, Saarbrucken)
Doug Arnold (University of Essex)
Ken Beesley (Xerox Research Centre Europe)
Gosse Bouma (University of Groningen)
Lynne Cahill (University of Sussex)
Bob Carpenter (Lucent Technologies)
Dick Crouch (University of Nottingham)
Michael Collins (University of Pennsylvania)
Morena Danieli (CSELT)
Ted E. Dunning (Aptex)
Michael Elhadad (Ben Gurion University)
Roger Evans (University of Brighton)
Dan Flickinger (Stanford University)
Gregory Greffenstette (Xerox Research Centre Europe)
Udo Hahn (University of Freiburgh)
Djoerd Hiemstra (University of Twente)
Barbora Hladka (Charles Univ., Prague)
David Israel (SRI International)
Arne Johnson (University of Linkoping)
Douglas A. Jones (Dept. of Defense)
Marcus Kracht (Free University of Berlin)
Geert-Jan Kruijff (Charles University, Prague)
Gina Levow (University of Maryland)
Marc Light (MITRE Corporation)
Inderjeet Mani (MITRE Corporation)
Andrei Mikheev (Harlequin)
Marc Moens (University of Edinburgh)
Christine Nakatani (Lucent Technologies)
Guenter Neumann (DFKI, Saarbruecken)
Gertjan van Noord (University of Groningen)
Karel Oliva (Univ. of Saarbruecken)
Richard Power (University of Brighton)
Steve Pulman (Cambridge University)
Ehud Reiter (University of Aberdeen)
Eileen Riloff (University of Utah)
Bill Rounds (University of Michigan)
Harold Somers (UMIST)
Tomasz Strzalkowski (General Electric)
John Tait (University of Sunderland)
Marc Villain (MITRE Corporation)
David Weir (University of Sussex)
Dekai Wu (Hong Kong Univ. of Science and Tech.)
Dan Zeman (Charles University, Prague)
Ingrid Zukerman (Monash University)
Steve Abney (AT&T Bell Labs)
Susan Armstrong (ISSCO, Univ. of Geneva)
Paolo Baggia (CSELT)
Bill Black (UMIST)
Eric Brill (CLSP Johns Hopkins Univ.)
Nicoletta Calzolari (CNR, Pisa)
Kenneth W. Church (AT&T Bell Labs)
Fabio Ciravegna (IRST)
Ann Copestake (Stanford University)
Judy Delin (University of Stirling)
Jan van Eijck (CWI)
Tomaz Erjavec (Ljubljana Univ.)
Frank van Eynde (Katholieke Universiteit Leuven)
Claire Gardent (University of Saarbrucken)
Ralph Grishman (New York University)
Jiri Hana (Charles Univ., Prague)
Julia Hirschberg (AT&T Bell Labs)
Beryl Hoffman (Marymount University)
Karen Jensen (Microsoft)
Mark Johnson (Brown University)
Ewan Klein (University of Edinburgh)
Hans-Ulrich Krieger (DFKI, Saarbruecken)
Ivana Kruijff (Charles University, Prague)
Hang Li (C&C Media Research Labs NEC Corp.)
Suresh Manandhar (University of York)
Mitch Marcus (Univ. of Pennsylvannia)
David Milward (SRI Cambridge)
Johanna Moore (University of Edinburgh)
Mark-Jan Nederhof (DFKI, Saarbruecken)
Anton Nijholt (University of Twente)
Kemal Oflazer (New Mexico State University)
Chris Paice (University of Lancaster)
Victor Poznanski (Sharp Laboratories of Europe)
Owen Rambow (CogenTex Inc)
Norbert Reithinger (DFKI, Saarbrucken)
Craige Roberts (Ohio State University)
Donia Scott (University of Brighton)
Richard Sproat (AT&T Bell Labs)
Marc Swerts (Institute for Perception Research)
Louis des Tombe (Utrecht University)
Juergen Wedekind (University of Stuttgart)
Janyce Wiebe (New Mexico State University)
Antonio Zampolli (CNR, Pisa)
Michael Zock (LIMSI)

Student Session
Preface

The student sessions have become an integral part of the EACL conferences. They provide an invaluable opportunity for young researchers to present their work to the community and receive feedback for future activities. In this spirit, we kept the tradition of encouraging students to submit not only papers presenting completed work (like for the main sessions), but also reports on work in progress.

We received 17 submissions from 8 countries, of which we accepted 8 papers for presentation and two as reserve papers. Each submission was reviewed at least by two student reviewers and one faculty reviewer.

We would like to thank the 31 reviewers for their fair assessment and their detailed comments, which we think were of great help for the student authors. Their names and affiliations are listed below.

We want to thank the Student Sessions Co-chairs of this year's ACL conference Melanie Baljko (Toronto, Canada) and Anna Korhonen (Cambridge, UK), and the Student Session Chairs of previous ACL and COLING-ACL conferences, in particular Pamela W. Jordan (University of Pittsburgh, USA), Maria Milosavljevic (CSRIRO, Australia), and Dragomir R. Radev (Columbia University, USA), for providing supporting material and for their cooperation. Finally, we want to thank the organizers of the main conference, more specifically Koenraad de Smedt (Bergen, Norway), Alex Lascarides (Edinburgh, UK), Giorgio Satta (Padova, Italy), and Henry Thompson (Edinburgh, UK), and the members of the European Chapter of ACL for their cooperation and help.

Jonas Kuhn, Atro Voutilainen

Programme Committee (Student Session)

Programme Committee (Poster Session)

Programme Chair:
Giorgio Satta (University of Padua)

Programme Committee Members:

Breck Baldwin (University of Pennsylvania)
Tilman Becker (DFKI, Saarbrücken)
Mari Broman Olsen (University of Maryland)
Federica Busa (Brandeis University)
Marie-Helene Candito (Universite Paris 7)
Thierry Declerck (DFKI, Saarbrücken)
Luca Dini (CELI, SNS, Italy)
Christy Doran (University of Brighton)
Jason Eisner (University of Pennsylvania)
Karin Harbusch (University of Koblenz)
Mark Hepple (University of Sheffield)
Elena Not (IRST, Italy)
Carolyn Penstein Rosé
Fabio Pianesi (IRST, Italy)
Emanuele Pianta (IRST, Italy)
Owen Rambow (CogenTex Inc.)
Norbert Reithinger (DFKI, Saarbrücken)
Srinivas Bangalore (AT&T Bell Labs)
Pasi Tapanainen (University of Helsinki)
Jakub Zavrel (Tilberg University)

Table of Contents

Student Session

Poster Session

Author Index

Student Author Index

Poster Author Index

EACL '99 Programme

June 9, 1999

0830 - 0930 Registration

0930 - 0940 Welcome

0940 - 1020 Main session 1

Auditorium 2

Andrei Mikheev, Marc Moens and Claire Grover: Named Entity Recognition without Gazetteers

Auditorium 3

Richard Power: Generating Referring Expressions with a Unification Grammar

1020 - 1100 Main session 2

Auditorium 2

Didier Bourigault and Christian Jacquemin: Term Extraction and Term Clustering: An Integrated Platform for Computer-Aided Terminology

Auditorium 3

Ian M. O'Neill and Michael McTear: An Object-Oriented Approach to the Design of Dialogue Management Functionality

1100 - 1130 Break

1130 - 1230 Invited talk

Auditorium 2

Bruce Croft: Language Models for Information Retrieval

1230 - 1400 Lunch

1400 - 1430 Student session 1

Auditorium 2

Gabriela Cavaglià: The Development of Lexical Resources for Information Extraction from Text Combining WordNet and Dewey Decimal Classification

Auditorium 3

Donna K. Byron and Joel R. Tetreault: A Flexible Architecture for Reference Resolution

1430 - 1510 Main session 3

Auditorium 2

Maria Lapata, Scott McDonald and Frank Keller: Determinants of Adjective-Noun Plausibility

Auditorium 3

Miriam Eckert and Michael Strube: Resolving Discourse Deictic Anaphora in Dialogues

1510 - 1550 Break

1550 - 1620 Student session 2

Auditorium 2

Patrick Caudal: Result States and the Lexicon: the Proper Treatment of Event Structure

Auditorium 3

Daniel S. Paiva: Investigating NLG Architectures: Taking Style into Consideration

1620 - 1700 Main session 4

Auditorium 2

Susanne Stevenson and Paola Merlo: Automatic Verb Classification Using Distributions of Grammatical Features

Auditorium 3

Pierre Boullier: Chinese Numbers, MIX, Scrambling and Range Concatentation Grammars

June 10, 1999

0900 - 0940 Main session 1
Auditorium 2
Glyn Morrill: Geometry of Lexico-Syntactic Interaction
Auditorium 3
Franz Josef Och: An Efficient Method for Determining Bilingual Word Classes

0940 - 1020 Main session 2
Auditorium 2
Inderjeet Mani, Therese Firmin, David House, Gary Klein, Beth Sundheim, Lynette Hirschman: The TIPSTER SUMMAC Text Summarization Evaluation
Auditorium 3
Tim Fernando: Ambiguous Propositions Typed

1020 - 1050 Break

1050 - 1150 Invited talk
Auditorium 2
Wolfgang Wahlster: Deep Processing of Shallow Structures: The Robust Integration of Speech, Language and Translation Technology for Intelligent Interface Agents

1150 - 1230 Main session 3
Auditorium 2
Rila Mandala, Takenobu Tokunaga and Hozumi Tanaka: Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval
Auditorium 3
Fabio Ciravegna and Alberto Lavelli: Full Text Parsing Using Cascades of Rules: An Information Extraction Perspective

1230 - 1400 Lunch

1400 - 1430 Student session 1
Auditorium 2
Justin Picard: Finding Content-bearing Terms using Term Similarities
Auditorium 3
Dimitrios Kokkinakis and Sophie Johansson Kokkinakis: A Cascaded Finite-State Parser for Syntactic Analysis of Swedish

1430 - 1510 Main session 4
Auditorium 2
Simone Teufel, Jean Carletta and Marc Moens: An Annotation Scheme for Discourse Level Argumentation in Research Articles
Auditorium 3
Thorsten Brants: Cascaded Markov Models

1510 - 1550 Break

1550 - 1620 Student session 2
Auditorium 2
Patrice Lopez: Repair Strategies for Lexicalized Tree Grammars
Auditorium 3
Veit Reuer: Dialogue Processing in a CALL-System

1620 - 1700 Main session 5
Auditorium 2
Dale Gerdemann and Gertjan van Noord: Transducers from Rewrite Rules with Backreferences
Auditorium 3
Giorgos Orphanos and Dimoitris Christodoulakis: POS Disambiguation and Unknown Word Guessing with Decision Trees

June 11, 1999

0900 - 0940 Main session 1
Auditorium 2
Maria Wolters and Mathias Kirsten: Exploring the Use of Linguistic Features in Domain and Genre Classification
Auditorium 3
Miguel Alonso, David Cabrero, Eric de la Clergerie and Manuel Vilares: Tabular Algorithms for TAG Parsing

0940 - 1020 Main session 2
Auditorium 2
Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis: Automatic Authorship Attribution
Auditorium 3
Guido Minnen: Selective Magic HPSG Parsing

1020 - 1050 Break

1050 - 1150 Poster and Demo session

1150 - 1230 Main session 3
Auditorium 2
Erik Tjong Kim Sang and Jorn Veenstra: Representing Text Chunks
Auditorium 3
Hiroyuki Shinnou: Detection of Japanese Homophone Errors by a Decision List Including a Written Word as a Default Evidence

1230 - 1400 Lunch

1400 - 1440 Main session 4
Auditorium 2
John Chen, Srinivas Bangalore and K. Vijay-Shanker: New Models for Improving Supertag Disambiguation
Auditorium 3
Kiyotaka Uchimoto, Satoshi Sekine and Hitoshi Isahara: Japanese Dependency Structure Analysis based on Maximum Entropy Models

1440 - 1520 Main session 5
Auditorium 2
Atro Voutilainen: An Experiment on the Upper Bound of Interjudge Agreement: The Case of Tagging
Auditorium 3
Fumiyo Fukumoto and Yoshimi Suzuki: Word Sense Disambiguation in Untagged Text Based on Term Weight Learning

1520 - 1535 Break

1535 - 1615 Business meeting (incl. TEI)
Auditorium 2

Posters

Izaskun Aldezabal, Inaki Alegria, Olatz Ansa, Jose Mari Arriola, Nerea Ezeiza, Itziar Aduriz and Alexander Da Costa: Designing Spelling Correctors for Inflected Languages Using Lexical Transducers

Hans Argenton and Anke Feldhaus: The Treegram Index—An Efficient Technique for Retrieval in Linguistic Treebanks

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin and John Tait: Simplifyng Text for Language-Impaired Readers

Nigel Collier, Hyun Seok Park, Norihiro Ogata, Yuka Tateishi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, Jun-ichi Tsuijii: The GENIA Project: Corpus-Based Knowledge Acquisition and Information Extraction from Genome Research Papers

Crit Cremers: A Note on Categorial Grammar, Disharmony and Permutation

Johann Gamper: Encoding a Parallel Corpus for Automatic Terminology Extraction

José Relaño Gil, Daniel Tapias, Maria C. Gancedo, Marcela Charfuelan and Luis A. Hernández: Robust and Flexible Mixed-Initiative Dialogue for Telephone Services

Adam Kilgariff: 95% Replicability for Manual Word Sense Tagging

Torbjörn Lager: μ-TBL: A Small, Extendible Transformation-Based Learner

Kaili Müürisep: Determination of Syntactic Functions in Estonian Constraint Grammar

John Nerbonne, Wilbert Heeringa and Peter Kieiweg: Comparison and Classification of Dialects

Frank Schilder: Pointing to Events

Mark Stevenson: A Corpus-Based Method for Deriving Lexical Mappings

Tutorials

The tutorials take place on June 8th, 1999.

Practical Text Mining 0930–1300
Lecturer: **Ronen Feldman** (Bar-Ilan University)

The information age has made it easy to store large amounts of data. The proliferation of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, while the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes. Text Mining is a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, NLP, IR and knowledge management. Text Mining involves the preprocessing of document collections (text categorization, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (distribution analysis, clustering, trend analysis, association rules etc) and visualization of the results. In this tutorial we will present the general theory of Text Mining and will demonstrate several systems that use these principles to enable interactive exploration of large textual collections. We will present a general architecture for text mining and will outline the algorithms and data structures behind the systems. Special emphasis will be given to efficient algorithms for very large document collections, tools for visualizing such document collections, the use of intelligent agents to perform text mining on the internet, and the use information extraction to better capture the major themes of the documents. The Tutorial will cover the state of the art in this rapidly growing area of research. Several real world applications of text mining will be presented.

Natural Language Learning with the Maximum Entropy Framework 1400–1730
Lecturer: **Adwait Ratnaparkhi** (IBM TJ Watson Research Center)

"Corpus-based" approaches to natural language processing (NLP), also known as "statistical" or "machine learning" approaches, have become popular in recent years due to the availability of large, annotated corpora. This tutorial will discuss how to implement a corpus-based NLP tool with the maximum entropy framework. We will first describe the maximum entropy framework, and then discuss its application to several problems, including sentence boundary detection, part-of-speech tagging, prepositional phrase attachment, parsing, and text categorization. Our experience has shown that this framework yields consistently high accuracies, requires relatively "knowledge-poor" informants, and is highly reusable across tasks. A general outline of the tutorial:

1. What is the maximum entropy framework ?
2. Using it for integrating diverse sources of evidence:
   - Sentence Boundary Detection
   - Part of Speech tagging
   - Parsing
3. Using it with and without annotated data:
   - Prepositional Phrase Attachment
4. Comparing its performance with other learning techniques:
   - Text Categorization
   - Incremental vs. Frequency-based feature selection
   - Decision Trees (C5.0)
5. Summary: Advantages and Disadvantages of the framework

Building Natural Language Generation Systems 0930–1300
Lecturers: **Robert Dale** (MRI, Macquarie University), **Ehud Reiter** (University of Aberdeen)

Natural language generation (NLG) systems produce understandable texts in English and other human languages from some underlying non-linguistic representation of information. NLG systems combine knowledge about language and the application domain to automatically produce documents, reports, explanations, help messages, and other kinds of texts.

In this tutorial we describe NLG from an applied system-building perspective; that is, we will explain how NLG systems are built. Our presentation will be based on a popular architectural model which encompasses the three stages of text planning, microplanning, and realisation. We will also give examples of current NLG applications; discuss when NLG technology is and is not appropriate; and explore how NLG can be integrated into multimedia, hypertext, and speech systems.

This tutorial should be useful for managers, implementors, and researchers. For managers, it will provide a broad overview of the field and what is possible today; for implementors, it will provide a realistic assessment of available techniques; and for researchers, it will highlight the issues that are important in current applied NLG projects.

Lexicography for Computationalists 1400–1730
Lecturers: **Adam Kilgarriff** (University of Brighton), **Michael Rundell**

NLP makes extensive use of dictionaries, but frequently proceeds in ignorance of lexicography. Even on computational lexicography projects, workers rarely have a background in lexicography. If workers in NLP were better informed about how dictionaries were produced, the purposes they were designed to serve, and what distinguished good dictionaries from bad ones, they would be better placed to choose a dictionary and to exploit the information it contained.

The tutorial will describe the goals and practices of corpus-based dictionary-making. Participants will gain an appreciation of the kinds of judgements that lexicographers need to make every day and the criteria they use. The tutorial will include practical exercises.

We shall also discuss models for lexicography/NLP collaboration, including SENSEVAL, the recent evaluation exercise for word sense disambiguation programs.